








Weakly supervised anomaly segmentation in retinal OCT images using an adversarial learning approach

JING WANG,^{1,2}  WANYUE LI,^{1,2}  YIWEI CHEN,²  WANGYI FANG,^{4,5} WEN KONG,^{1,2} YI HE,^{2,6}  AND GUOHUA SHI^{2,3,7} 

¹ Jiangsu Key Laboratory of Medical Optics, Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Science, Suzhou 215163, China

² School of Biomedical Engineering (Suzhou), Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230026, China

³ Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Science, Shanghai 200031, China

⁴ Department of Ophthalmology and Vision Science, Eye and ENT Hospital, Fudan University, Shanghai 201112, China

⁵ Key Laboratory of Myopia of State Health Ministry, and Key Laboratory of Visual Impairment and Restoration of Shanghai, Shanghai 200003, China

⁶ heyi@sibet.ac.cn

⁷ ghshi_lab@126.com

Abstract: Lesion detection is a critical component of disease diagnosis, but the manual segmentation of lesions in medical images is time-consuming and experience-demanding. These issues have recently been addressed through deep learning models. However, most of the existing algorithms were developed using supervised training, which requires time-intensive manual labeling and prevents the model from detecting unaware lesions. As such, this study proposes a weakly supervised learning network based on CycleGAN for lesions segmentation in full-width optical coherence tomography (OCT) images. The model was trained to reconstruct underlying normal anatomic structures from abnormal input images, then the lesions can be detected by calculating the difference between the input and output images. A customized network architecture and a multi-scale similarity perceptual reconstruction loss were used to extend the CycleGAN model to transfer between objects exhibiting shape deformations. The proposed technique was validated using an open-source retinal OCT image dataset. Image-level anomaly detection and pixel-level lesion detection results were assessed using area-under-curve (AUC) and the Dice similarity coefficient, producing results of 96.94% and 0.8239, respectively, higher than all comparative methods. The average test time required to generate a single full-width image was 0.039 s, which is shorter than that reported in recent studies. These results indicate that our model can accurately detect and segment retinopathy lesions in real-time, without the need for supervised labeling. And we hope this method will be helpful to accelerate the clinical diagnosis process and reduce the misdiagnosis rate.

© 2021 Optical Society of America under the terms of the [OSA Open Access Publishing Agreement](#)

1. Introduction

Lesion provides a gold standard for initial disease diagnosis and subsequent treatment. Lesion identification and localization are thus central objectives for common imaging modalities such as magnetic resonance (MR), computed tomography (CT), and optical coherence tomography (OCT). The huge number of medical images, for example, there are approximately 30 million OCT procedures performed to detect retinopathy worldwide each year [1], provides extensive sample data for research purposes. However, these images also make it labor-intensive and time-consuming for doctors to review them manually. Some traditional methods have been

proposed for the auxiliary reading of medical images, including level set [2] and kernel regression [3] techniques. But these algorithms are typically slow, insufficiently robust, and overly sensitive to noise. Deep learning (DL), which has been heavily studied in medical imaging, can often overcome these issues. DL algorithms can generally be divided into supervised and unsupervised categories, depending on whether labels are used in the training process.

Deep supervised learning models, which require elaborate labeling, have been proposed for lesion segmentation in annotated medical images. Kamnitsas et al. proposed a 3D CNN for segmenting brain lesions in MRIs [4]. Chen et al. proposed a dense-res-inception net to segment multi-lesion structures in CT and MR brain images [5]. Lesion segmentation has also been extensively studied in OCT images. For instance, Hu et al. proposed a deep neural network with spatial pyramid pooling for segmenting subretinal fluid (SRF) and pigment epithelium detachment (PED) lesions [6]. Ruwan et al. proposed an Unet-based architecture consisting of encoding and de-coding blocks with skip-connections to segment intraretinal fluid (IRF), SRF, and PED in labeled retinal OCT images [7]. Rhona et al. developed a novel multi-decoder framework to segment drusen in OCT scans [8]. Each of these algorithms performed well with custom or public datasets.

However, the training of supervised segmentation models needs large quantities of annotated images with pixel-level labels, which requires diagnostic expertise and can be time-consuming and cost-prohibitive. Besides, some annotations may lack sufficient detail for specific applications, which will result in mislabeled or omitted subtle lesions that limit prediction capacity.

Unsupervised learning has attracted increased attention recently as it can be used without labels. For example, Tajayasu et al. proposed a two-phase approach using joint unsupervised learning and k-means clustering for pathological segmentation of lung cancer in micro-CT images [9]. Chen et al. implemented an active contour without edges framework via a convolutional neural network (CNN) to achieve high-quality bone segmentation in single-photon emission computed tomography (SPECT) images [10]. Each of these models implemented unsupervised segmentation using feature selection and clustering, which is sensitive to outliers and requires significant computational runtime.

Between fully supervised learning and unsupervised learning, there is weakly supervised learning, where the learning model can be trained with incomplete, inexact, or inaccurate labels. It mitigates the need for full labels and makes sure the model is learning what we want. For example, Hoel et al. proposed a weakly supervised model for cardiac image segmentation, they only used 0.1% ground-truth labels but reached a performance close to full supervision [11]. And it's typical for training a pixel-level lesion segmentation model with image or volume level labels in the medical images, for instance, Wang et al. trained a classification model on chest CT images to detect COVID-19 infectious and they located lesions by detecting the activation regions of the model [12], similarly, Ma et al. proposed to segment geographic atrophy (GA) in retinal OCT images by calculating the class activation map from a trained GA classification model [13].

The present study is comparable to recent works applying weakly supervised learning by image translation. For example, Philipp et al. trained an autoencoder on healthy retinal OCT images with image-level labels and used a one-class support vector machine to identify anomalies in new data [14]. Similarly, Thomas et al. used a GAN-based technique to train a generative model by labeled healthy retinal OCT scans. This algorithm successfully detected abnormalities in new data with a combined anomaly score based on the trained model [15]. These and other studies have successfully detected lesions by applying an autoencoder or GAN to identify normal pathology, distinguishing abnormal markers by evaluating the posterior probability of test samples generated by the trained model. However, these models were never introduced to real abnormal samples during training, it's difficult for them to guarantee the output is paired with the correct input in the testing stage. Especially, as variations widely exist in retinal shapes and spatial orientations in OCT images, it is tough to blindly acquire matching positive and negative retina

samples. In state-of-the-art, preprocessing steps like layer segmentation, flattening, and patch clipping are often included to reduce the impact of unpaired OCT data and match unassigned images when testing, and additional post-processing steps are required to concatenate patches into full-width-images [16]. Such steps can be highly time-consuming and difficult to optimize or automate, which limits the clinical application of these techniques.

CycleGAN [17], which implements a ‘cycle consistent loss’ concept to achieve unpaired image-to-image translation, has been widely used for unpaired style transfer in medical imaging. This has included stain normalization in multi-center whole slide images (WSIs) [18], style transfer between different lung X-ray datasets [19], and image variability reduction between retinal images acquired with different OCT devices [20]. Moreover, some studies have proved that CycleGAN can be implemented to detect lesions in brain MRI images and histology imagery, where the appearance of lesions won’t change the shape of the anatomy severely [21–22]. The results of these studies demonstrate the potential of the CycleGAN algorithm for transferring texture styles in unpaired images.

To the best of our knowledge, few methods of image translation have been applied in the context of weakly supervised lesion segmentation in full-width retinal OCT images due to the variations in shape, thickness, and spatial orientation of the retinas. In this paper, a novel technique, based on the CycleGAN algorithm, is proposed to detect and segment lesions in full-width retinal OCT images by image translation. We trained a generative model to ‘repair’ the deformed anatomy structures of the input abnormal samples by generating paired normal samples (it can be explained as that the generated sample looks like the treated abnormal sample, where only lesions area were reconstructed and others remain the same as the original sample). Then the lesions markers were segmented by a simple comparison between the input and output images. The whole segmentation process was blind to the ground-truth of the pixel-level segmentation map, only image-level labels are needed, which is much easier to acquire. The framework for this model is shown in Fig. 1.

CycleGANs were initially designed to transfer complex local textures between image domains (i.e., CT and MR), not necessarily between objects of different shapes. But retinal lesions (such as SRF, IRF) often lead to changes in retina shape, original CycleGAN didn’t perform well on this mission. In this study, a customized network, where a dilated convolutional block-based discriminator combined with a U-net generator, and a multi-scale structural similarity perceptual reconstruction loss (MS-SSIM) were used to solve this problem. The customized architecture can capture more global structure variabilities and the MS-SSIM can represent geometric differences using area statistics, they can further assist the model in focusing on structural changes [23]. This approach was implemented using a public retinal OCT dataset, provided by [32], which included more than 100,000 images with retinopathy labels. Results demonstrated that our model performed well in reconstructing lesion areas from abnormal to normal anatomy, even in samples with large shape deformations.

The main contributions of this paper can be summed as:

- 1) We proposed a CycleGAN based model to generate normal-look full-width retinal OCT image from the input abnormal image and achieved lesion segmentation by calculating the difference of them.
- 2) Instead of going through the complex pre-process like retina flattening and region of interest clipping, the input images are fed to the model directly. A dilated convolutional block-based discriminator and an MS-SSIM loss are introduced to overcome the variations in shape, thickness, and spatial orientation of the retinas.
- 3) The proposed model works well in reconstructing normal-look retinal OCT images, even when the retina is heavily distorted by the lesions. And a post-process is implemented to locate multiple lesions.

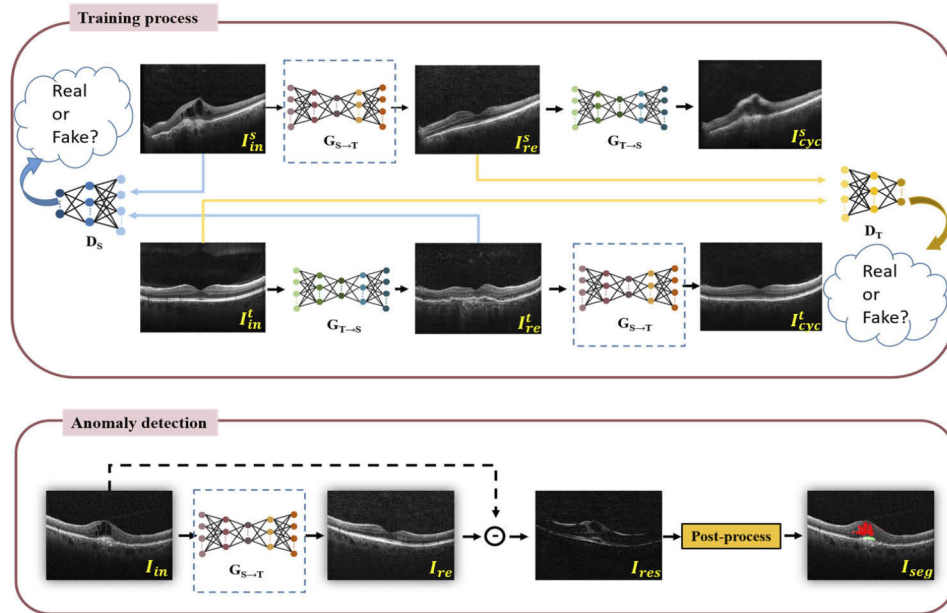


Fig. 1. The training process (top) and the anomaly detection process (bottom) for the proposed model. After training, the $G_{S \rightarrow T}$ is able to reconstruct normal samples (I_{re}) from an abnormal input (I_{in}). Then the difference between the input and output is calculated (I_{res}), and a further post-process is implemented to locate lesions (I_{seg}).

2. Method

The proposed technique was developed using a CycleGAN architecture, as shown in Fig. 1, and typically consisted of two generators ($G_{S \rightarrow T}$, $G_{T \rightarrow S}$) and two discriminators (D_S , D_T). The $G_{S \rightarrow T}$ is trained to generate target domain image from source domain input, to deceive the D_T , who is trained in an adversarial manner to distinguish fake target domain images generated by $G_{S \rightarrow T}$ from real target domain images (and vice versa). In this study, the model was trained using retinal OCT images, which were separated into abnormal (S) and normal (T) categories and input to $G_{S \rightarrow T}$ and $G_{T \rightarrow S}$ for synchronous training. During testing, annotated as anomaly detection in Fig. 1, the positive samples (I_{in}) are input to $G_{S \rightarrow T}$ to generate corresponding negative samples (I_{re}), and the difference of the input and output is calculated (I_{res}). Next, a post-processing step is implemented to identify the appearance of lesion markers (I_{seg}). Finally, the fluid and exudation areas were highlighted in the input OCT images. The architecture of the generator and discriminator are introduced in section 2.1, network training details are discussed in section 2.2, a description of the objective function is provided in section 2.3, and lesion identification post-processing is explained in section 2.4.

2.1. Generator and discriminator

The generator implemented in the original CycleGAN model was a residual block-based convolutional network, in which residual blocks were only applied at a single scale (3×3) in deep layers [17]. This residual generator was capable of extracting deep, semantic, and coarse-grained feature maps. However, shallow, low-level, and fine-grained feature maps (mostly containing edge, contour, and location information) were left out. As our primary objective is to translate retinal images of varying shapes, this single-scale residual block is problematic, since it prevents the generator from acquiring sufficient structural or location information. As such, a U-net shape

architecture was used to construct the generators in our study [24], by combining deep and shallow feature maps to provide a more precise output implementing skip connections, as shown in Fig. 2(a). The network is in symmetric architecture, the left part is an Encoder that can extract features from the input images and the right part is a Decoder that can construct the output images from the extracted features. The Encoder was built in a fully convolutional network formation. It concludes eight convolutional blocks, where each contains a 4×4 convolutional layer with a stride of 2 and followed by a leaky rectified linear unit (Relu). Similarly, the Decoder consisted of eight transposed convolutional blocks, where each contains a 4×4 upconvolutional layer with a stride of 2 and followed by a leaky ReLU, except the last block, which was activated by a tanh function. We use the 4×4 kernel size for widening the receptive field. This U-net structure helped the generator to acquire multi-scale features and overcome the effects of shape or location variability.

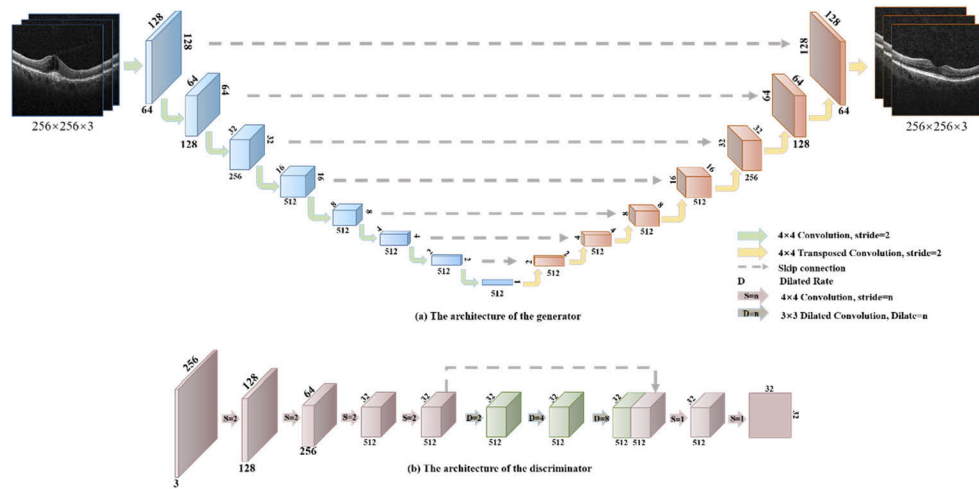


Fig. 2. The (a) generator and (b) discriminator architectures.

The CycleGAN implemented PatchGANs [25–27], initially developed for common GAN applications such as style transfer and texture synthesis, as the discriminators. However, the PatchGAN determines real or fake scores by evaluating image patches of a fixed size, which prevents the network from perceiving global spatial information and causes the generator to perform poorly for objects of varying shapes. Thus, a dilated convolutional network was used here to resolve this issue [28], which can widen the receptive field of the network by incorporating data from a larger region without introducing parameters, the network architecture is shown in Fig. 2(b). The discriminator contains six 4×4 convolutional layers and three 3×3 dilated convolutional layers, where all layers with corresponding strides (1 or 2) and followed by a Leak ReLU, except the last layer, which without activation function. And the dilated convolutional layers with dilated rates of 1, 2, and 4, respectively. It determines real/fake images by finding real/fake regions in the images from a larger surrounding instead of judging a fixed size local image patch, which helps the generator focus on those abnormal regions better.

2.2. Network training

Prior to training the anomaly detection model, input data were manually separated into normal and abnormal categories based on retinal anatomy. Unlike conventional weakly supervised algorithms, which typically train using only the normal set, the proposed model was trained using both data categories. Retinal images containing abnormal and normal anatomy were annotated as

x_{sn} and x_{Tn} , respectively, then input to $G_{S \rightarrow T}$ and $G_{T \rightarrow S}$. As the retinal images differ widely in appearance and retinopathies often exhibit shape changes or retinal thickening, making it difficult to synthesize paired images. Previous studies on disease marker recognition in OCT images [14–15] have relied on extensive preprocessing, including outer layer segmentation, flattening, and patch clipping to adjust for variations in orientation, shape, or sample thickness, which might change the lesions appearances or sizes, and it's difficult for clinicians to visualize the lesions in patches. In the present study, this issue was solved by applying the cycle consistent loss to learn the bijective mapping between the two image domains and the self-supervised synthesis process [17]. No preprocessing steps were applied to the training data except image-level labeling, the model was trained to generate full-width paired normal images from the full-width abnormal input OCT images. The reconstructed retinas were located at nearly the same position as in the input images and only lesion areas were reconstructed into the normal anatomy.

Model performance was assessed using a set of images y_{sm} not seen during the training process. This test set included corresponding pixel-level binary segmentation maps $S_m \in \{0, 1\}$, where fluid-filled areas were labeled with '1' and other areas were labeled with '0'. Image-level labels $l_m \in \{0, 1\}$ were also added, with normal and abnormal images labeled as '0' and '1', respectively. Test images then input to $G_{S \rightarrow T}$ to generate paired sets. Finally, the image backgrounds of the original images and reconstructed images, which are often complex and noisy, were removed by retina edge detection during pixel-level lesion segmentation.

2.3. Objective function

The intended purpose of the CycleGAN model is to transfer images of different styles, wherein the appearance of objects remains mostly unchanged. The corresponding CycleGAN objective function can be expressed as:

$$L(G_{S \rightarrow T}, G_{T \rightarrow S}) = L_{GAN}(G_{S \rightarrow T}) + L_{GAN}(G_{T \rightarrow S}) + \eta_1 L_{cyc} + \eta_2 L_{identity}, \quad (1)$$

where L_{GAN} represents the GAN loss, η_1 and η_2 are balance parameters. Cycle consistency loss L_{cyc} was used to constrain translations to be reversible and identity mapping loss $L_{identity}$ was used to generate more realistic images, details can be found in [17].

In this study, normal retinal anatomy was reconstructed from abnormal images, in which most retinopathy structures exhibited variations in shape. Excepting the customized network architecture, the Multi-scale structure similarity perceptual reconstruction loss (MS-SSIM) [29] was also used to compensate for this variability. The SSIM is a perceptual derived metric to assess the structural similarity between two images simulate humans, lots of works have proved that the SSIM performs better on assessing image quality than mean-based methods [30,31]. It can be formulated as follows:

$$\begin{aligned} SSIM(x, y) &= l(x, y) * c(x, y) * s(x, y) \\ &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} + \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} + \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \\ &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} + \frac{\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \end{aligned} \quad (2)$$

where C_1, C_2 are constants and $\mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{xy}$ denote means, standard deviations, and cross-covariance of the image pair (x, y) from G and the corresponding input image respectively. MS-SSIM is the multiscale extension of the SSIM, which can be formulated as follow:

$$MS_SSIM(x, y) = \prod_j^M SSIM(x_j, y_j), \quad (3)$$

where (x_j, y_j) is the j^{th} image patch and M is the scale level. It is more flexible than SSIM.

The MS-SSIM can recognize geometric differences using area statistics, but often overlooks smaller details and does not capture color similarity. As such, combining MS-SSIM with L1 or L2 loss, which can calculate pixel-level differences, can provide a more optimal representation. In this study, cycle consistency loss L_{cyc} in the initial CycleGAN was replaced by MS-SSIM loss (L_{ss_cyc}). The corresponding objective function can be expressed as:

$$L(G_{S \rightarrow T}, G_{T \rightarrow S}) = L_{GAN}(G_{S \rightarrow T}) + L_{GAN}(G_{T \rightarrow S}) + \eta_1 L_{ss_cyc} + \eta_2 L_{identity}, \quad (4)$$

the L_{ss_cyc} is calculated by:

$$L_{ss_cyc} = \lambda_{ss} L_{ss} + \lambda_{l1} L_1, \quad (5)$$

$$L_{ss} = (1 - \text{MS_SSIM}(rec_S, real_S)) + (1 - \text{MS_SSIM}(rec_T, real_T)), \quad (6)$$

$$L_1 = l_1(rec_S, real_S) + l_1(rec_T, real_T), \quad (7)$$

where reconstructed source domain images $rec_S = G_{T \rightarrow S}(G_{S \rightarrow T}(x))$, reconstructed target domain images $rec_T = G_{S \rightarrow T}(G_{T \rightarrow S}(x))$, l_1 represents the mean absolute error, MS represents MS-SSIM metric, and λ_{ss} , λ_{l1} , η_1 and η_2 are balance coefficients.

2.4. Anomaly detection

An anomaly score was also implemented to quantify deviations between abnormal and paired reconstructed normal retinal images [15]. The metric used in this study can be expressed as:

$$A(x) = \|x - G_{S \rightarrow T}(x)\|^2 + \|f(x) - f(G_{S \rightarrow T}(x))\|^2, \quad (8)$$

where f is the feature layer before the final layer in D_T . This anomaly score should be lower for normal-looking images and higher for anomalous images. Since $G_{S \rightarrow T}$ was only trained to generate normal images, $G_{S \rightarrow T}(x)$ was visually similar to normal retina images, regardless of the value of input x (which can be normal samples or abnormal samples).

In addition to acquiring image-level classification results, the following metric was also used to calculate pixel-level differences:

$$\dot{A}(x) = x - G_{S \rightarrow T}(x). \quad (9)$$

Lesion localization is typically conducted by directly comparing the $\dot{A}(x)$ with a threshold. However, retinal OCT scans often exhibit structural variations or thickening that prevent the use of a single threshold for every anomaly type. In this study, as shown in Fig. 3, abnormality location was separated into the following steps: (1) as there is complicate random background noise in the input OCT images, which is hard for the model to mimic perfectly, we segment the top layer (ILM) and the bottom layer (RPE) of the retina to remove the background of the input and output images with the graph search based edge detection algorithm [33] firstly (Fig. 3(a) and Fig. 3(b)). (2) Then the residual image was acquired following Eq. (9) ((c) in Fig. 3, the image shows the $\dot{A}(x)$). (3) To better observe the residual pixels, we applied automatic binarization to the residual image by the OTSU algorithm [34] (in Fig. 3(d)). (4) A mask was generated from the residual image with the supervision of the edges and separated the residual image into two parts: the overlap (marked by yellow in Fig. 3(e)) and non-overlap (marked by red in Fig. 3(e)). (5) For the overlap, if the pixel value of the residual image $\dot{A}(x) > 0$ and the binary image $B(x) = 255$, the pixel was labeled as exudates (labeled as green in Fig. 3(f)), and if the $\dot{A}(x) < 0$, the $B(x) = 255$, the pixel is labeled as fluid (labeled as yellow in Fig. 3(f)). For the non-overlap, where $B(x) = 0$ was labeled as fluid (labeled as red in Fig. 3(f)). Finally, we concatenate all the detection results to generate the whole segmentation map ((g) in Fig. 3, where exudate was labeled by green, fluid was labeled by red).

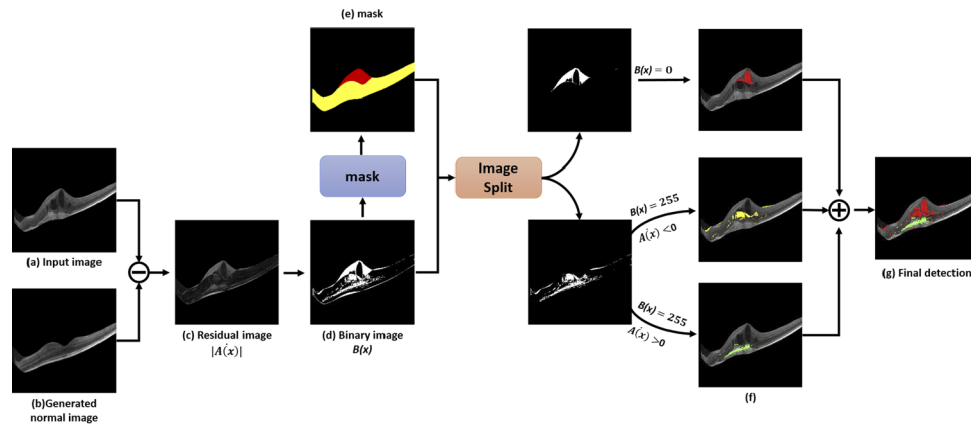


Fig. 3. The post-processing steps involved in anomaly detection. Subtracting the input image (a) and the output image (b) to get the residual image (c). The binarized image (d) of (c) is separated into two parts according to the mask (e), and lesions are detected in overlap (highlight by yellow) and non-overlap (highlight by red) according to the pixels values. The final segmentation map is (g).

3. Experiment

The proposed model was trained and evaluated using a publicly available retinal OCT image dataset, obtained from [32]. Another public dataset containing binary segmentation maps, obtained from [3], was also implemented to assess the model's robustness. These two image groups are hereafter denoted the K's dataset and Chiu's dataset, respectively. The model was evaluated by determining (1) if the generated images were realistic, (2) if the model could detect abnormal retina images, and (3) if the lesions could be accurately located. The proposed method was compared with three existing algorithms, the f-AnoGAN [15], CycleGAN [17], and Ganimorph [23] models. Ablation experiments were further implemented to determine the effects of various network architectures and loss strategies.

3.1. Data

The proposed network was trained using K's dataset, a large labeled retinal OCT image dataset with more than 100,000 images, which was acquired using the Spectralis OCT system (Heidelberg Engineering, Germany) from 5,319 patients [32]. These images included training, testing, and validation data, which were annotated into four categories: diabetic macular edema (DME), choroidal neovascularization (CNV), drusen, and normal. This dataset was initially constructed for retinopathy classification and includes augmented images that have been rotated, tailored, resized, or has been added random noise (which are commonly used image augmentation methods), these augmentation images were initially implemented to prevent the overfitting problem, whose appearances are far from the real images and we found they are limited effective in improving the models' performance but caused a huge increase of the training time (some examples of the augmented images can be found in Supplement 1). So we excluded those images with severe appearance changes. A total of 12,765 and 8,891 images were acquired from the 'normal' and 'CNV' categories, respectively, as few qualified images were available in the other categories. These selected images were set as the training set of this work.

Unlike many conventional weakly supervised image generation algorithms, our method does not require complex preprocessing steps prior to training, as the selected full-width images were directly inputted to the network. The K's data contains an independent test set with 250 images in each retinopathy category. We selected all images in the CNV and the normal categories to test

our model. And the DME test images were also selected to test the robustness of the model, as lesion shapes of the DME differed significantly from the CNV, and the DME lesion is completely unknown to the model. Finally, 500 abnormal images (250 CNV, 250 DME) and 250 normal images were set as the test set of our work. The K's dataset did not include pixel-level anomaly labels, two trained retinopathy reviewers with more than three years of experience provided labels for fluid-filled regions in the testing images, and a clinical physician reviewed the result and fixed the wrong labels.

Network robustness was further tested using Chiu's dataset, which included retinal OCT images from ten DME patients [3]. The data were acquired using the Spectralis OCT system (Heidelberg Engineering, Germany) and featured a resolution of 768×496. This set included 78 manually labeled images, displaying corresponding fluid area segmentation results. The labeled images were used to test the model and provided a comparison with baseline results acquired by the kernel regression method, reported in [3].

All images were scaled into a resolution of 256×256 to fit the model, and the segmentation maps were also resized to the same resolution. Test images were inputted directly to the network to generate paired images. However, the OCT data included significant levels of background noise, which impeded accurate lesion detection. In the final lesion segmentation step, an automated graph search-based edge detection algorithm [33] was used to segment the top and bottom layers of the retina to remove the background.

3.2. Training and evaluation details

The f-AnoGAN [15], CycleGAN [17], and Ganimorph [23] networks were also trained using the K's dataset, to provide a comparison with the proposed model. All models processed 256×256 input images and were trained for 40 epochs using two NVIDIA 2080Ti GPUs with a batch size of 2. As f-AnoGAN was initially designed for generating images with a 64×64 resolution, two additional layers were added to the generator and discriminator, to produce images with a 256×256 resolution. Two test sets, the K's data and Chiu's data with pixel-level segmentation labels, were constructed to evaluate the trained network. And we empirically set the hyperparameters as $\lambda_{ss} = 0.5$, $\lambda_{l1} = 0.5$, $\eta_1 = 1$, $\eta_2 = 0.5$.

Qualitative evaluation: Results were assessed visually as images were presented to two trained OCT image readers with more than three years of experience. These participants evaluated a Turing test set [28], consisting of 50 real normal retinal OCT images and 50 synthetic images, in an attempt to differentiate generated and real data. The synthetic images were reconstructed by the trained $G_{S \rightarrow T}$ from normal (9 images) and abnormal samples (21 CNV images and 20 DME images). Input data exhibited a resolution of 256×256 and were acquired from the K's test dataset. The two readers provided classification results independently.

Quantitative evaluation: The proposed model was also evaluated quantitatively using two test sets with both image-level and pixel-level labels, to assess anomaly detection accuracy. The Image-level classification results were acquired by computing the anomaly score stated in Eq. (8). Classification results were compared with three related algorithms, f-AnoGAN [15], CycleGAN [17], and Ganimorph [23]. The f-AnoGAN model achieved unsupervised anomaly detection in OCT images with a GAN-based technique, but they didn't get an accurate lesion segmentation map. CycleGAN is a basic algorithm used for unpaired image texture transfer. Ganimorph is an improved version of CycleGAN, designed to be compatible with transfer between objects of varying shapes. The pixel-level anomaly detection results were acquired following Fig. 3.

3.3. Results

Qualitative results: The qualitative results can be found in Fig. 4. The samples of DME, CNV, and negative were fed to the model to generate respective normal-like images. The residual images were generated by subtracting the generated image from the input image to better observe the

difference between the input and output images. We found that the CycleGAN failed to reconstruct normal sample from the DME sample and tends to cause some artifacts in the reconstruction of the CNV samples, but it performs well in the normal samples. For the Ganimorph, it tends to produce some artifacts in the generated images of the positive samples as the red arrows show in Fig. 4, and it tends to produce some unnecessary changes in the reconstruction of the negative sample, as shown in the residual images. The f-AnoGAN can generate normal samples from the input, however, it has not only changed the abnormal areas but also changed the normal areas, the location of the retina, and the background. And its generated images are not realistic. Reconstructing the abnormal areas in the positive samples without severe artifact or extra change and keeping the reconstructed sample the same as the input negative sample, the proposed method possessed the best performance in the reconstruction of positive and negative samples.

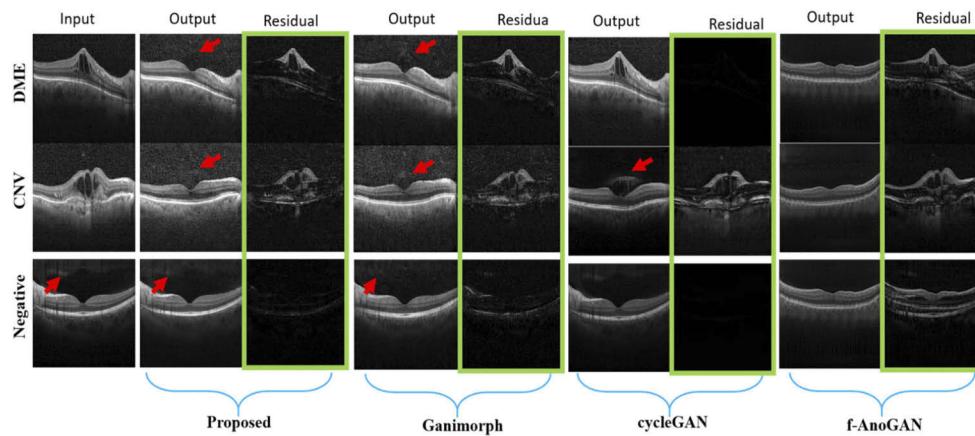


Fig. 4. Qualitative results produced by the proposed algorithm. The images in the first row are real images, including normal (columns 1), CNV (columns 2), and DME (columns 3) types. The second row shows the corresponding generated normal-like images.

Besides, a Turing test was conducted using two trained OCT image readers who qualitatively evaluated the generated results of the proposed method, as discussed above. The accuracy of differentiating generated images from real images was 14% (7 images were recognized from a set of 50 generated images and 4 real images were misclassified as synthetic images) and 16% (8 generated images were recognized and 4 real images were misclassified) for the two readers, respectively. The image readers were provided with images generated from DME samples, CNV samples, and negative samples in the test set, as shown in Fig. 4. It is evident the trained model performed well in representing normal anatomical variability and transferring between objects with shape deformations, even for anomalies unseen during the training process. This was evidenced by the DME samples, which were not included in the training set.

Quantitative results: Image-level anomaly detection accuracy was compared with three comparative algorithms, f-AnoGAN [15], CycleGAN [17], and Ganimorph [23]. These results are presented in Table 1, with the highest values in bold. The corresponding receiver operating characteristic (ROC) curve, the area under the curve (AUC), and precision-recall (PR) scores are provided in Fig. 5. These results suggest that our proposed method outperformed comparable models in image-level anomaly detection. Our model can also generate normal anatomy retinal images from abnormal retina images in an average of 0.039 seconds, which is significantly shorter than the patch-based methods.

Lesion segmentation: Pixel-level anomaly detection was performed on the K's data (including CNV and DME samples) and Chiu's data, and the fluid and exudates were detected. This technique was also compared with lesion segmentation results produced by the CycleGAN and Ganimorph

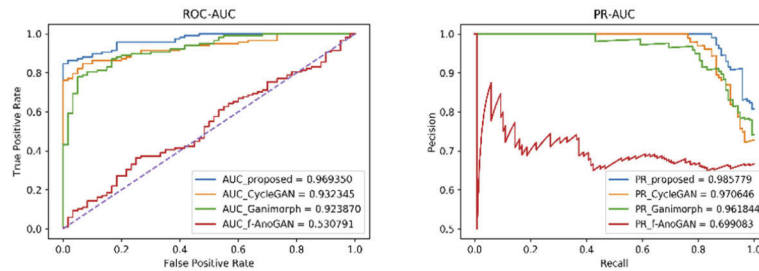


Fig. 5. The ROC curve (left) and precision-recall curve (right) and corresponding AUC scores of the proposed technique, CycleGAN, and Ganimorph.

Table 1. Quantitative results using the proposed technique, f-AnoGAN [15], CycleGAN [17], and Ganimorph [23]. Precision, sensitivity, F1-score, specificity, and AUC score were calculated to evaluate Image-level anomaly detection performance

| | f-AnoGAN | CycleGAN | Ganimorph | Proposed |
|-------------|----------|----------|-----------|---------------|
| Precision | 0.6982 | 0.9341 | 0.9321 | 0.9671 |
| Sensitivity | 0.5021 | 0.8025 | 0.7897 | 0.8374 |
| F1-score | 0.5841 | 0.8633 | 0.8550 | 0.8976 |
| Specificity | 0.4976 | 0.7315 | 0.738 | 0.7864 |
| AUC | 0.5308 | 0.9323 | 0.9239 | 0.9694 |

algorithms. Results from the K's and Chiu's datasets are shown in Fig. 6. The f-AnoGAN algorithm was excluded from the pixel-level segmentation experiment, as in this experiment setting, training images were fed to the model directly, which results in the generated images of the f-AnoGAN differing significantly from the input and with serious artifacts, as shown in Fig. 4, the pixel-level results could not be acquired without the inclusion of pre-processing, like flatten and clip. More examples can be found in the [Supplement 1](#).

In K's dataset, the Ganimorph works well in reconstructing CNV samples into normal (as the cyan box shows in Fig. 6) but encounters problems in reconstructing the DME samples, which with more serious shape deformation (as the orange box shows in Fig. 6). It can't translate some abnormal structures into normal anatomy (as the yellow arrows show in Fig. 6), which affects the edge detection and so the lesion location results. The CycleGAN works worse than the Ganimorph, it failed in translating samples in the DME and CNV categories, especially for the lesion areas with shape deformation.

In Chiu's dataset (as the blue box shows in Fig. 6), the Ganimorph can generate normal samples from the input images, but there are still apparent artifacts in the output images as the yellow arrow shows in Fig. 6. The CycleGAN failed to reconstruct samples in this dataset. The proposed method achieves superior performance in these two data sets, reconstructing lesion areas into normal anatomy and preserving other normal areas. Furthermore, the proposed method and the CycleGAN works better in preserving the background of the input image while the Ganimorph doesn't, as the green arrows show in Fig. 6.

These results indicate the proposed model outperformed comparable methods in generating plausible normal anatomy images from abnormal data with shape deformations and locating lesions. The CycleGAN model produced better results than Ganimorph in image-level anomaly detection but exhibited the worst performance in generating realistic normal images and lesion segmentation. In general, CycleGAN worked well for transferring texture styles but failed to translate examples with evident shape deformations. Ganimorph is better at overcoming the shape deformation and reconstructing normal-look images but its outputting images often lack

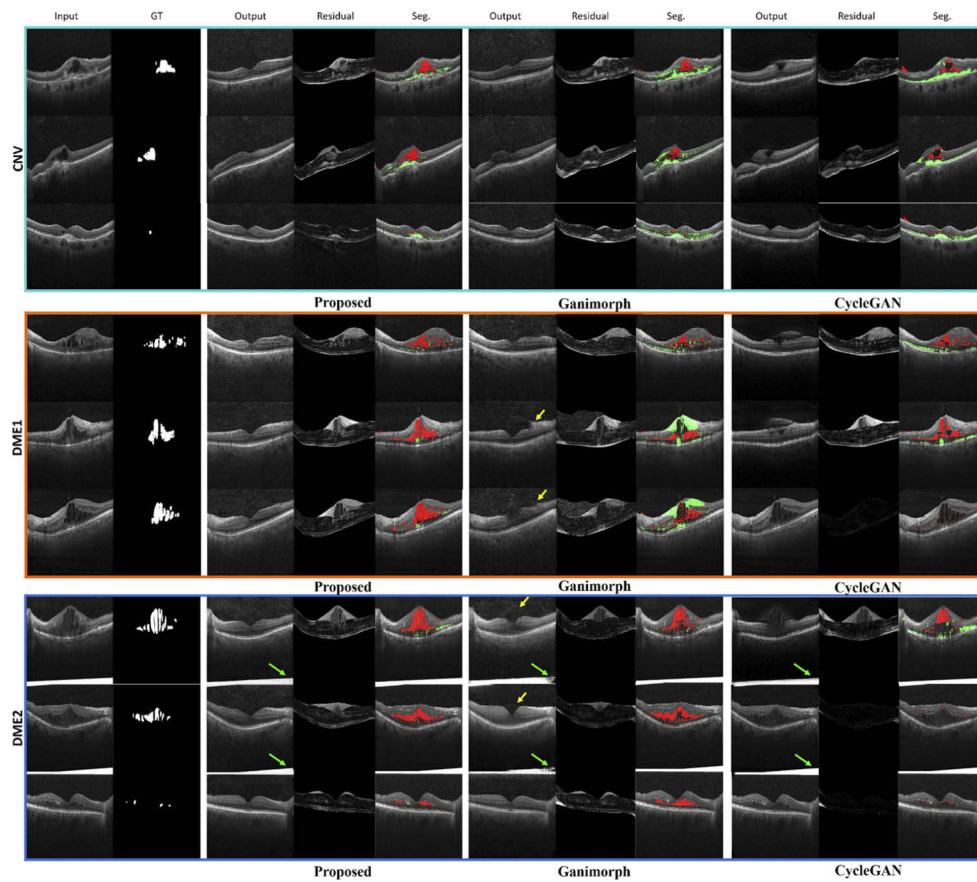


Fig. 6. Pixel-level anomaly detection results for the CNV and DME samples in K's dataset (marked with the cyan and orange boxes, respectively), and DME samples in the Chiu's dataset (marked with the blue box). The input images, the ground truth of the segmentation map (annotated as GT in the figure), the residual images, and the segmentation map (annotated as Seg. in the figure) are provided. As illustrated, the Ganimorph generates artifacts (the yellow arrows) and failed to preserve the background (the green arrows).

certain details and contain apparent artifacts. And images generated with Ganimorph were noisier and more blurry than images produced using the proposed method, as demonstrated in Fig. 6. More qualitative results can be found in the [Supplement 1](#).

To better observe the capability of the model, we also measured the Dice coefficient for fluid segmentation. All test samples with fluid were selected, the segmentation maps generated by the mentioned models were compared with the ground truth. We reported the mean Dice coefficients as shown in Table 2. The proposed method achieves the best performance in both two datasets, which is commensurate with the qualitative results, indicating that the proposed model achieved better lesion detection capabilities than alternative methods. This was particularly evident with Chiu's dataset, in which our fluid segmentation Dice value (0.64) was higher than the value (0.53) reported by Chiu et al. [3], where a kernel regression method was implemented.

Ablation experiments: An ablation experiment was used to determine the contribution of MS-SSIM and the architecture of the discriminator and generator to the result. The resulting images and segmentation maps are shown in Fig. 7 and the corresponding mean Dice coefficients for the fluid segmentation are provided in Table 3. We first implemented the original CycleGAN

Table 2. Dice coefficients for lesion segmentation of the K's dataset and Chiu's dataset, acquired using CycleGAN [17], Ganimorph [21], and the proposed model. The baseline was acquired from the original study by Chiu et al. [3] and the best results are highlighted in bold.

| | K's Dataset | Chiu's Dataset |
|----------------|---------------|----------------|
| Baseline [3] | — | 0.53 [3] |
| CycleGAN [17] | 0.3785 | 0.2527 |
| Ganimorph [23] | 0.7025 | 0.4590 |
| Proposed | 0.8239 | 0.6444 |

architecture to reconstruct corresponding normal images from the input images, the model is good at translating the texture but failed to handle the shape deformation, as shown in Fig. 7(1). Then we replaced the cycle consistency loss L_{cyc} in the CycleGAN with the proposed (formulated as Eq. (7)), referred to as 'CycleGAN + SS'. The model devoted attention to structural variability but the quality of the results suffered for images exhibiting large shape deformations, as shown in Fig. 7(2). We then replaced the original Resnet block-based generator with the proposed U-net generator (referred to as 'CycleGAN + SS + U_Ge'). The results in Table 3 demonstrate that the Dice coefficients improved heavily with this architecture. However, the model still performed poorly in reconstructing retinas with severe anatomical warping, since the original patch-based discriminator cannot capture global spatial information, as shown in Fig. 7(3). Then we replaced the original patch-based discriminator with the dilated convolution-based discriminator and removed the MS-SSIM metric (referred to as 'CycleGAN + U_Ge + Di_Dis'). The result demonstrates that the lesion areas were captured and corresponding normal anatomic structures were generated in most cases, but there were still some artifacts in the generated images, as shown in Fig. 7(4). Finally, we added the MS-SSIM metric to overcome these problems (referred to as 'CycleGAN + SS + U_Ge + Di_Dis'), as it can better preserve the perceptual features instead of noisy high-frequency information, this approach resulted in the most normal-looking images, as seen in Fig. 7(5). In addition, dilated discriminator was combined with resnet block-based generator (denoted 'CycleGAN + SS + Res_Ge + Di_Dis'), which led to the mode collapse displayed in Fig. 7(6). This was in part because the resnet block-based generator at a single scale limits the discriminator from capturing sophisticated features.

Table 3. Dice coefficients for fluid segmentation of the ablation experiments. The loss objects, discriminator architecture, and generator architecture were replaced to investigate the effects of varying components.

| Model | Discriminator | | Generator | | Loss | | Dice Coefficients | |
|---|---------------|---------|-----------|------|-----------|--------------|-------------------|---------------|
| | Patch | Dilated | Resnet | Unet | L_{cyc} | L_{ss-cyc} | K's Data | Chiu's Data |
| CycleGAN | √ | | √ | | √ | | 0.3785 | 0.2527 |
| CycleGAN + SS | √ | | √ | | | √ | 0.4951 | 0.4103 |
| CycleGAN + SS + U_Ge | √ | | | √ | | √ | 0.6830 | 0.4458 |
| CycleGAN + U_Ge + Di_Dis | | √ | | √ | √ | | 0.7014 | 0.3644 |
| CycleGAN + SS + U_Ge + Di_Dis (Proposed) | | √ | | √ | | √ | 0.8239 | 0.6444 |
| CycleGAN + SS + Res_Ge + Di_Dis | | √ | √ | | | √ | Failed | Failed |

Table 3 shows the mean Dice coefficients for fluid segmentation produced by different network architectures. The combination of MS-SSIM, U-net generator, and the dilated discriminator produced the highest coefficients of 0.8239 and 0.6444 in the K's dataset and the Chiu's data, respectively. The quantitative results are commensurate with the qualitative results, both of them have proved the superiority of the proposed method. And these results suggested that the

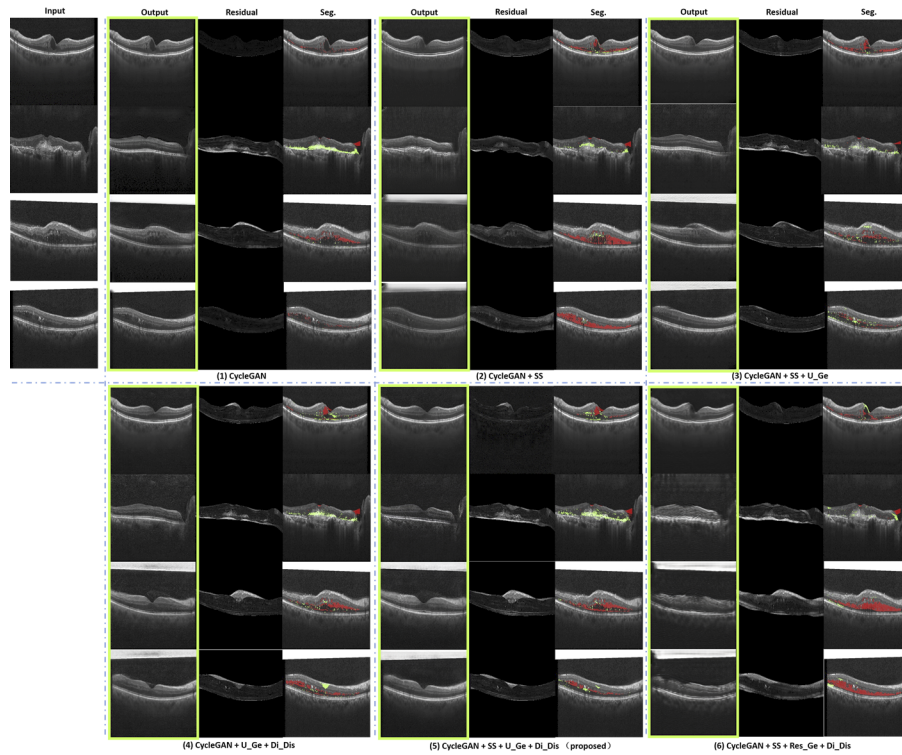


Fig. 7. Qualitative results of the ablation experiments. The same input images were fed to all mentioned models. The output image (marked by the green boxes), residual images and images with segmentation map of the (1) CycleGAN, (2) 'CycleGAN + SS', (3) 'CycleGAN + SS + U_Ge', (4) 'CycleGAN + U_Ge + Di_Dis', (5) 'CycleGAN + SS + U_Ge + Di_Dis', and the (6) 'CycleGAN + SS + Res_Ge + Di_Dis' are provided.

original CycleGAN structure works well for texture transfer but struggles with shape deformations. However, including MS-SSIM can help the model focus on structural inconsistencies and regions containing pixel variations. The U-net generator can learn multi-scale features used to reconstruct realistic texture. A dilated discriminator can help the model capture global context information and transfer the abnormal retina to a corresponding normal shape.

4. Conclusion and discussion

In this paper, a new methodology is presented for weakly supervised anomaly segmentation of retinal OCT images. This technique achieved anomaly segmentation by subtracting generated normal-looking anatomy from corresponding input abnormal retina images, as is common in unsupervised anomaly detection. However, unlike conventional algorithms (most of which use a single GAN model), where a complex multi-step preprocess was implemented to reduce the effects of shape deformation and location variability, the proposed model uses a CycleGAN-based network architecture to permit training with unpaired images. The model was trained using full-width original OCT data and only implemented background removal in the final step of pixel-level lesion segmentation, it saves much time and is user-friendly (the appearances of lesion and retina are unchanged).

The cycleGAN was initially proposed to transfer texture between different image domains but is not ideal for transferring between images exhibiting shape deformations. As such, the patch

discriminator and Resnet block-based generator were replaced with a dilated discriminator and a U-net generator, respectively. Multi-scale structure similarity perceptual reconstruction loss was also included to help the model adapt to this unique transfer task. The network was trained and tested using subsets of a public K's database and Chiu's dataset. The proposed model got 96.64% AUC and 0.8239 dice coefficient on public K's dataset, outperformed comparable algorithms in both image-level anomaly screening and pixel-level lesion segmentation. On the other public dataset, the model gets a 0.64 Dice coefficient, which is 0.11 higher than that in the original study.

It is also worth noting that our method achieved transformations between full-width retina OCT images with an average time of 0.039 s, which is significantly faster than the patch-based methods. In conclusion, we have demonstrated that our proposed technique can achieve real-time style transfer for images exhibiting structural variations. It is also capable of accurate pixel-level anomaly segmentation and should be generally applicable to unsupervised lesion contouring in other unpaired medical data, particularly valuable for images whose anatomic structure might vary due to the presence of lesions.

However, some issues remain to be solved. First, since the training set consists only of images through the macular, so the model can't generate correct images when the input images are around the macular. This problem can be resolved by adding a few retinal images of non-macular areas to the training set. Second, we found that the background noise of OCT images is difficult to mimic due to randomness, especially when the lesion caused a huge shape deformation to the retina. In this method, we have implemented a graph search-based edge detection method to reduce the effect of the background noise, but this method is easy to affect by lesions. So a better method of background noise reduction remains essential. Third, we found in the results that the model was unable to translate the images with extremely severe lesions, in which the anatomical structures were nearly indistinguishable, examples can be found in the [Supplement 1](#). We are still working on finding a way to detect lesions in such images.

Funding. National Key Research and Development Program of China (2016YFF0102002); National Natural Science Foundation of China (61605210, 62075235); Youth Innovation Promotion Association of the Chinese Academy of Sciences (2019320); Jiangsu Provincial Key Research and Development Program (BE2019682); Entrepreneurship and innovation talents in Jiangsu Province (Innovation of scientific research institutes).

Acknowledgments. The authors would like to thank Prof. Zhang Kang, Prof. Sina Farsiu, and their groups for providing the public data sets.

Disclosures. The authors declare that there are no conflicts of interest related to this article.

Data availability. Data underlying the results presented in this paper are available in [\[30\]](#).

Supplemental document. See [Supplement 1](#) for supporting content.

References

1. E. A. Swanson and J. G. Fujimoto, "The ecosystem that powered the translation of OCT from fundamental research to clinical and commercial impact," *Biomed. Opt. Express* **8**(3), 1638–1664 (2017).
2. J. Novosel, K. A. Vermeer, J. H. de Jong, Z. Wang, and L. J. van Vliet, "Joint segmentation of retinal layers and focal lesions in 3-D OCT data of topologically disrupted retinas," *IEEE Trans. Med. Imaging* **36**(6), 1276–1286 (2017).
3. S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," *Biomed. Opt. Express* **6**(4), 1172–1194 (2015).
4. K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.* **36**, 61–78 (2017).
5. L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "DRINet for medical image segmentation," *IEEE Trans. Med. Imaging* **37**(11), 2453–2462 (2018).
6. J. Hu, Y. Chen, and Z. Yi, "Automated segmentation of macular edema in OCT using deep neural networks," *Med. Image Anal.* **55**, 216–227 (2019).
7. R. Tennakoon, A. K. Gostar, R. Hoseinnezhad, and A. Bab-Hadiashar, "Retinal fluid segmentation in OCT images using adversarial loss based convolutional neural networks," Presented at 2018 *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. [Online].

8. R. Asgari, J. I. Orlando, S. Waldstein, F. Schlanitz, M. Baratsits, U. Schmidt-Erfurth, and H. Bogunović, "Multiclass segmentation as multitask learning for Drusen segmentation in retinal optical coherence tomography," *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019).
9. T. Moriya, H. R. Roth, S. Nakamura, H. Oda, K. Nagara, M. Oda, and K. Mori, "Unsupervised segmentation of 3D medical images based on clustering and deep representation learning," Presented at *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging* (2018). [Online].
10. J. Chen and E. C. Frey, "Medical image segmentation via unsupervised convolutional neural network," Presented at *Medical Imaging with Deep Learning 2020* (2020, January).
11. H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. Ben Ayed, "Constrained-CNN losses for weakly supervised segmentation," *Med. Image Anal.* **54**, 88–99 (2019).
12. X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, and C. Zheng, "A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT," *IEEE Trans. Med. Imaging* **39**(8), 2615–2625 (2020).
13. X. Ma, Z. Ji, S. Niu, T. Leng, D. L. Rubin, and Q. Chen, "MS-CAM: Multi-scale class activation maps for weakly-supervised segmentation of geographic atrophy lesions in SD-OCT images," *IEEE J. Biomed. Health Inform.* **24**(12), 3443–3455 (2020).
14. P. Seeböck, S. M. Waldstein, S. Klimesch, H. Bogunovic, T. Schlegl, B. S. Gerendas, and G. Langs, "Unsupervised identification of disease marker candidates in retinal OCT imaging data," *IEEE Trans. Med. Imaging* **38**(4), 1037–1047 (2019).
15. T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.* **54**, 30–44 (2019).
16. Y. He, A. Carass, Y. Liu, B. M. Jedynek, S. D. Solomon, S. Saidha, and J. L. Prince, "Deep learning based topology guaranteed surface and MME segmentation of multiple sclerosis subjects from retinal OCT," *Biomed. Opt. Express* **10**(10), 5042–5058 (2019).
17. J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," Presented at *Proceedings of the IEEE international Conference on Computer Vision* (2017).
18. T. de Bel, M. Hermesen, J. Kers, J. van der Laak, and G. J. Litjens, "Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology," Presented at *MIDL* (2019).
19. C. Chen, Q. Dou, H. Chen, and P. A. Heng, "Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation," Presented at *International workshop on machine learning in medical imaging*. (2018, September).
20. D. Romo-Bucheli, P. Seeböck, J. I. Orlando, B. S. Gerendas, S. M. Waldstein, U. Schmidt-Erfurth, and H. Bogunović, "Reducing image variability across OCT devices with unsupervised unpaired learning for improved segmentation of retina," *Biomed. Opt. Express* **11**(1), 346–363 (2020).
21. C. Baur, R. Graf, B. Wiestler, S. Albarqouni, and N. Navab, "SteGANomaly: inhibiting CycleGAN steganography for unsupervised anomaly detection in brain MRI," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, eds. (Springer International Publishing, 2020), pp. 718–727.
22. D. Stepec and Skocaj, "Unsupervised detection of cancerous regions in histology imagery using image-to-image translation," *arXiv preprint arXiv:2104.13786* (2021).
23. A. Gokaslan, V. Ramanujan, D. Ritchie, K. In Kim, and J. Tompkin, "Improving shape deformation in unsupervised image-to-image translation," Presented at *Proceedings of the European Conference on Computer Vision (ECCV)* (2018, September).
24. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *In International Conference on Medical image Computing and Computer-assisted Intervention* (Springer, 2015), (pp. 234–241).
25. P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," Presented at *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017) [Online].
26. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, and A. Acosta, W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4681–4690.
27. C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," Presented at *European Conference on Computer Vision* (2016).
28. F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *In: International Conference on Learning Representations* (2015).
29. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Process.* **13**(4), 600–612 (2004).
30. C. You, Q. Yang, H. Shan, L. Gjestebj, G. Li, S. Ju, and G. Wang, "Structurally-sensitive multi-scale deep neural network for low-dose CT denoising," *IEEE Access* **6**, 41839–41855 (2018)..
31. L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Trans. on Image Process.* **20**(8), 2378–2386 (2011).
32. D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, and K. Zhang, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell* **172**(5), 1122–1131 (2018).

33. M. K. Garvin, M. D. Abramoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka, "Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images," *IEEE Trans. Med. Imaging* **28**(9), 1436–1447 (2009).
34. N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.* **9**(1), 62–66 (1979).